PATENT APPLICATION
DOCKET NO.: 1285-0123US
ALC-139133

"EXPRESS MAIL" Mailing Label No..EV331251378US.....
Date of Deposit.....AUGUST 28, 2003.....

# DISTRIBUTED AND DISJOINT FORWARDING AND ROUTING SYSTEM AND METHOD

## BACKGROUND OF THE INVENTION

### Technical Field of the Invention

**[0001]** The present invention generally relates to communications networks. More particularly, and not by way of any limitation, the present invention is directed to an architecture for implementing a distributed and disjoint forwarding and routing scheme that provides for high availability in a communications network.

### Description of Related Art

**[0002]** Core IP backbone networks are evolving to support more than just Internet data traffic. Much of the traffic that now needs to run over IP - such as voice, video, and mission-critical business traffic through virtual private networks (VPNs) - requires higher availability than the conventional data applications (e.g., email, web browsing, et cetera) demand. To support the growing portion of the Internet

communications containing mission-critical traffic, core routers will need to be improved so they can provide the high availability now required.

**[0003]** However, today's currently deployed routers are not designed to provide this new level of availability that the new IP traffic requires, e.g., 99.999% uptime (commonly referred to as "five nines" availability). In essence, they lack the robustness and key features that would allow carriers to achieve five nines availability - one of which being the ability to maintain forwarding and routing during failures and upgrades.

## SUMMARY OF THE INVENTION

**[0004]** Accordingly, the present invention advantageously provides a distributed and disjoint forwarding and routing system and method operable with a routing element having a scalable cluster-based architecture, wherein the control plane and data plane are loosely-coupled for effectuating non-disruptive switchover in the event of a failure.

**[0005]** In one aspect, the present invention is directed to a router that includes a partitionable data plane having one or more forwarding tables and a partitionable control plane having one or more routing tables operating under control of at least one routing protocol process. Each forwarding table includes a forwarding information base (FIB) operable to effectuate

a forwarding process through the router with respect to incoming data, e.g., packets, cells, frames, etc. Each routing table includes a routing information base (RIB) related to the applicable routing protocol for effectuating routing decisions with respect to the data forwarding process through the router. A control plane update agent module is provided for maintaining a redundant set of routing table information in at least one control plane update buffer, wherein the control plane update agent module is operable to synchronize the routing tables in the control plane in a time-based or event-based manner, or both. A data plane update agent module is operably coupled to the control plane update agent module for asynchronously coordinating the updating of the forwarding table information based on the routing table information in association with a set of data plane update buffers. In the event of a failure, the data forwarding process continues to proceed based on information stored in at least one of the data plane or control plane update buffers even as a switchover operation is underway in the router.

**[0006]** In one embodiment, the data plane and control plane update agent modules may be integrated into a single inter-plane updating mechanism disposed between the data and control planes for mediating the updating and coordination process therebetween. In another embodiment, the data and control planes may be logically partitioned into a plurality of virtual partitions, each

with one or more data plane nodes and one or more control plane nodes, respectively. The data plane nodal complex and the control plane nodal complex may each be organized into a separate scalable cluster-based network having any known or heretofore unknown topology, e.g., a topology selected from the group consisting of ring topologies, star topologies, Clos topologies, toroid topologies, hypercube topologies, or polyhedron topologies, to name a few. By way of an exemplary implementation, a data plane node may include one or more processing engines, one or more forwarding tables with associated update buffers and a data plane update agent. Likewise, a control plane node may include one or more control processors, one or more routing tables with associated update buffers and a control plane update agent.

**[0007]** In another aspect, the present invention is directed to a fault-tolerant routing element having a distributed scalable architecture. A logic structure, e.g., a management module with process status monitoring (PSM) capability, which may comprise any means implemented in software, hardware, firmware, or in combination thereof, is provided for detecting a fault in an active node disposed in the routing element that is engaged in executing a router process. Another structure is provided for effectuating a continuous switchover from the active node to a redundant node responsive to detecting a fatal fault, whereby the redundant node continues to execute the router process without

disruption in data forwarding. An updating means is provided for updating routing table information and forwarding table information associated with the routing element responsive to the switchover operation.

**[0008]** In a still further embodiment, the present invention is directed to a distributed network wherein the capability of continuous switchover is effectuated by loosely-coupling the control and data planes over the network. The distributed network comprises at least a first network element operable to route data responsive to applicable control messages provided thereto. At least a second network element is operably coupled to the first network element, wherein the network elements are comprised of a router with decoupled and disjoint control and data planes.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0009]** A more complete understanding of the present invention may be had by reference to the following Detailed Description when taken in conjunction with the accompanying drawings wherein:

**[0010]** FIG. 1 depicts a high level functional block diagrammatic view of a fault-tolerant routing element having a distributed and disjoint forwarding and routing system and method in accordance with an embodiment of the present invention;

**[0011]** FIG. 2A depicts a functional block diagram of a data plane (DP) node disposed as part of a scalable cluster of a plurality of nodes of the routing element shown in FIG. 1;

**[0012]** FIG. 2B depicts a functional block diagram of a control plane (CP) node disposed as part of a scalable cluster of a plurality of nodes of the routing element shown in FIG. 1;

**[0013]** FIG. 3 depicts an embodiment of the invention with partitionable data and control planes with an update agent layer disposed therebetween;

**[0014]** FIG. 4 depicts an exemplary implementation of a router in accordance with the teachings of the present invention; and

**[0015]** FIG. 5 depicts a flow chart of the various operations involved in effectuating a fault-tolerant non-disruptive routing methodology in accordance with an embodiment of the present invention.

## DETAILED DESCRIPTION OF THE DRAWINGS

**[0016]** In the drawings, like or similar elements are designated with identical reference numerals throughout the several views thereof, and the various elements depicted are not necessarily drawn to scale. Referring now in particular to FIG. 1, depicted therein is a high level functional block diagrammatic view of a fault-tolerant routing element 100 having a distributed and

disjoint forwarding and routing system and method in accordance with an embodiment of the present invention. It should be recognized by those skilled in the art that the routing element 100 is operable to interwork within any type of communications network, organized in known or heretofore unknown network architecture, for effectuating routing functionality with respect to a variety of communications, e.g, data, voice-over-IP, video-over-IP, and other mission-critical applications. Further, the routing element 100 is deemed to be compatible with a host of routing protocols for routing all types of data (packets, cells, and the like, for instance.)

**[0017]** As can be readily seen, the structure and functionality of the routing element 100 may be logically segregated into two planes, a control plane (CP) 102A and a data plane (DP) 102B, that are loosely-coupled for effectuating routing decision-making functionality and data forwarding functionality, respectively. Accordingly, CP 102A and DP 102B may also be referred to as "routing plane" and "forwarding plane," respectively. Each of the planes may preferably be embodied as a cluster-based, scalable distributed network, partitionable into one or more nodes. Reference numeral 104A refers to a plurality (N) of CP nodes interconnected in any known or heretofore unknown network topology, e.g., a topology selected from the group comprised of ring topologies, star topologies, Clos topologies, toroid topologies, hypercube topologies, or polyhedron

topologies, just to name a few. Likewise, reference numeral 104B refers to M instances of DP nodes interconnected in a particular topology.

**[0018]** As will be described in greater detail below, each CP or DP node is responsible of effectuating routing functionality or data forwarding functionality, respectively, as part of the network cluster within which it is disposed. Further, CP 102A is operable to receive control inputs or updates (collectively, control messages) from other network elements (e.g., peer routers) to which the routing element 100 is coupled. By way of example, these control messages may comprise router status/availability messages, compatibility messages, routing protocol-specific messages, et cetera. Similarly, CP 102A is capable of generating appropriate control messages towards other peer elements in the communications network.

**[0019]** Although not specifically depicted in FIG. 1, it should be understood that each of the planes is provided with appropriate databases and processing elements (e.g., routing tables with routing information base or RIB instances in the CP domain and forwarding tables with forwarding information base or FIB instances in the DP domain) that may be distributed among the cluster nodes of the plane. In order to achieve fault tolerance in the routing element 100, an inter-plane updating mechanism 106 is disposed between the CP cluster 104A and the DP cluster 104B for providing a loose

coupling therebetween, whereby conventional synchronization process between the CP databases (i.e., RIB tables) and DP databases (i.e., FIB tables) is disjoined so that database updating is performed asynchronously. Consequently, as will be described below in further detail, the effect of a fault-induced switchover on the database updating process is not propagated from one plane to the other, thereby greatly reducing the need to arrest or suspend the operation of the entire routing element.

**[0020]** FIG. 2A depicts a functional block diagram of a DP node 200 disposed as part of a scalable cluster of a plurality of nodes of the routing element 100 shown in FIG. 1. Node 200 may comprise a logical node or a physical node, and may be representative of a data path node disposed in the DP cluster 104B. As a data path node, it can be a logical node in the sense of a partition or a part of the data plane, or a physical implementation, e.g., a line card in the ingress/egress of a port of the switching fabric of the routing element 100. In general, the functionality of node 200 involves responding to data traffic input and populating a database table 202 (e.g., a FIB) that is customized for the node. A local update buffer 206 is provided for maintaining a copy of the node-specific database under control of an update agent 204. An external signaling and control block 208 is provided for transmitting and

receiving update messages to and from other data plane nodes via path 210.

**[0021]** By way of example, as a data plane node, node 200 includes a forwarding agent that is operable to switch the incoming packets or cells to appropriate output port of the logical/physical node based on the entry of the node-specific FIB that may be redundantly provided as database 202 and its corresponding update buffer 206. The update agent 204 is operable to update the entries in the FIB, mediated via the update buffer 206 when it is needed and requested and remains dormant otherwise. In turn, the update buffer 206 is coordinated with a corresponding RIB update buffer residing in the CP by means of a CP update agent as exemplified by CP node 200B shown in FIG. 2B. In other words, the data and control paths of the routing element are rendered disjoint, whereby the RIB and FIB tables are maintained, updated, and redundantly engineered, independently of the state of one another. Accordingly, when a fatal software/hardware failure is encountered in the control path (either due to intentionally-caused maintenance downtime or because of a legitimate fault), the CP domain may undergo a switchover from an active node to a standby node, but the forwarding agents in the DP domain may continue to execute packet forwarding processes for existing paths based on current update buffer conditions.

**[0022]** Referring to FIG. 2B in particular, CP node 200B includes several functional and structural entities

that are similar in their overall architectural interrelationship to the entities described above with respect to DP node 200A. As a control path node, node 200B is representative of the nodes disposed in the CP cluster 104A, and in general, it may also be provisioned as a logical partition. Responsive to control updates received from other peer entities in the network, CP node 200B is operable to maintain and update routing information with respect to the traffic encountered in the network. Further, responsive to internal control inputs and upstream control messages, CP node 200B is operable to generate appropriate control messages towards other peer elements in the network. A database 250 is provided for maintaining an RIB for the network element. A local update buffer 254 is provided for maintaining a copy of the RIB, whereby a node-specific update agent 252 is operable in conjunction with the RIB database(s) and buffer(s) for synchronizing with other RIB tables and their copies in the network element, mediated via a control and external signaling block 254 that is coupled to other CP nodes via path 256 when the control plane is active. In addition, the node-specific update agent 252 is operable to coordinate the updating of the network element's forwarding tables in an asynchronous manner using the node-specific DP update agents. A management module 258, which may be provided as part of a CP node or otherwise associated therewith, is operable to interact in the CP domain for providing appropriate management

control with respect to the various CP nodes. For instance, management control may include forming groups and/or sub-groups of management nodes and selecting single or multiple group "leaders" by an election algorithm (e.g., a round-robin scheme). Due to redundant provisioning, when a management module (or a set) fails for some reason, another module or set may take over the control or leadership. Additional functionalities relating to management control in the context of effectuating continuous switchover in a network element will be set forth below.

**[0023]** Since the updating process between the CP domain nodes and DP domain nodes is coordinated and mediated via node-specific update agents, it should be appreciated that the FIB information in the DP nodes may not be reflective of the entire RIB of a network element but may be indicative of only what needs to be forwarded by a particular DP. Accordingly, the asynchronous updating process between the CP and DP domains results in partial updating of the forwarding data. Only specific information relevant to a DP node (e.g., a line card) is on the line card, and when a line card fails, the management/control plane decides which other line cards may take up the redistributed load (based on load balancing criteria, for example). Appropriate information may then be sent to the line cards via the update agents for redistributing the loads accordingly.

**[0024]** Referring now to FIG. 3, shown therein is an embodiment of the invention with partitionable data and control planes 303, 307, respectively, with an update agent plane 305 disposed therebetween. In one implementation, the partitionable data and control planes may be resident in a scalable router 300 that is architected as a plurality of partitions, virtual, logical or physical, 302-1 through 302-K. Each partition may be comprised of one or more DP nodes and one or more CP nodes, with a portion of the update agent functionality coupling both types of nodes. By way of illustration, each partition is shown to include one DP-Node and one CP-Node, each node having its update agent that is operable to communicate with the other update agent, or with the agents of other partitions (in the same planar domain or across the plane boundary). Reference numerals 304-1 through 304-K refer to the K data path nodes that correspond to the control nodes 306-1 through 306-K.

**[0025]** The partitions of the control plane may be organized into multiple CP blades with redundancy, where a separate instance of each control plane process can run on each blade, one of the blades being active and the other blades being standby. Regardless of the blade architecture, each CP node of the partition includes an update agent that controls coordination between a routing table (i.e., RIB) used for effectuating the routing process supported by the node and its update buffer

image. Although not specifically shown, one or more control processors are included for executing a routing process application based on applicable protocols. Since the router 300 may be implemented in a variety of communications networks for operation in conjunction with diverse peer elements, failover protection may be provided for several protocol processes executing on different nodes, e.g., Border Gateway Protocol (BGP), Intermediate System to Intermediate System (IS-IS), Open Shortest Path First (OSPF), Generalized Multi-Protocol Label Switching (GMPLS), Internet Group Management Protocol (IGMP), and the like.

**[0026]** Continuing to refer to FIG. 3, reference numerals 316-1 through 316-K, 318-1 through 318-K, and 314-1 through 314-K refer to the K routing tables, associated update buffers and CP update agents, respectively. A management/administration functionality structure may be provided as a separate module or integrated within one of the CP nodes for monitoring and detecting faults throughout the system.

**[0027]** Each data path node includes a node-specific and protocol-specific forwarding table (i.e., FIB) that is updated by a DP update agent in association with an update buffer. As explained in the foregoing discussion, the DP update agent is loosely coupled to a CP update agent for updating the entries of the FIB maintained at the data path node. Reference numerals 308-1 through 308-K, 310-1 through 310-K, and 312-1 through 312-K refer

to the K forwarding tables, associated update buffers and DP update agents, respectively. Further, reference numerals 320-1 through 320-K, 322-1 through 322-K, and 324-1 through 324-K refer to the various coupling paths associated with CP-DP update agent coupling, cross-partition update agent coupling, and inter-DP update agent coupling.

**[0028]** In addition, those skilled in the art should recognize that although the control and data planes 303 and 307, and update agent plane 305 have been particularly described in reference to a scalable router architecture, the functionality of the disjointed planes may be distributed over a number of network elements interconnected in a communications network. Thus, a distributed network may be advantageously provisioned with elements having continuous switchover capability in accordance with the teachings of present invention. By way of exemplary implementation, such a distributed network can be symmetric, i.e. involving elements that have similar or the same capacity, or asymmetric (with nodes having different capacities). Moreover, such a distributed network may be topologically symmetric (i.e., regular) or asymmetric (i.e., irregular), as well.

**[0029]** FIG. 4 depicts an exemplary implementation of a router 400 in accordance with the teachings of the present invention. One or more CP cards, e.g., CP cards 402-1 and 402-2, may be organized into redundant CP blades, with one blade being active and the other(s)

being switchover standby blade(s). Each CP card, which includes a routing process (RP) engine or module and associated database(s), may be inter-coupled to the other portions of the system, e.g., other CP cards, line cards, et cetera, via a system bus 416. Reference numerals 404-1 and 404-2 refer to two RP engines associated with the two exemplary CP cards 402-1, 404-2, respectively. Likewise, two database modules 406-1 and 406-2 are exemplified with respect to the CP cards for maintaining redundant RIB instances. An update agent 414 is provided to effectuate protocol stack synchronization among the various CP cards.

**[0030]** Reference numerals 420-1 and 420-2 refer to a plurality of line card partitions of the router 400, wherein each line card is operable to support a number of ports e.g., ports 426-1 and 426-2, which can be optical, electrical, or opto-electrical ports. A data forwarding process (FP) engine or module (reference numerals 422-1 and 422-2) and associated database(s) (reference numerals 424-1 and 424-2) are provided for each line card for effectuating Layer-2 packet forwarding operations. A data path update agent 418 is operably coupled to the control path update agent 414 and the various line card partitions for updating and synchronizing local FIB(s) with other nodes and CP database(s) only when the CP is not down, i.e., at least one CP blade is active.

**[0031]** A management module 408 including a process status monitoring (PSM) block 412 is provided for

effectuating management control with respect to fault tolerance. The PSM block 412 is operable to monitor various processes and modules of the router 400 for status changes so that conditions that may indicate software or hardware failures may be detected. Further, the PSM's functionality may also include determining whether a failure comprises a fatal error or fault that necessitates a continuous switchover (i.e., maintaining data forwarding processes non-disruptively while the router internally transitions to standby processes and modules). Additional related functionality of the PSM block 412 may include facilitating exchange of "heartbeat" messages between the processes on the active and standby CP nodes, and establishing internal consistency checks that monitor hardware and software applications.

**[0032]** Referring now to FIG. 5, depicted therein is a flow chart of the various operations involved in effectuating a fault-tolerant routing methodology in accordance with an embodiment of the present invention. In a monitoring operation, a router process status is checked (block 502) for status changes, internal consistencies, and the like, in an architecture that support loosely-coupled synchronization between its CP and DP nodes mediated by means of asynchronous and disjoint CP and DP update agent processes (block 504). Accordingly, as described hereinabove, there is no tight coupling between the various routing processes, i.e.,

each process is supported independently of another. Further, the updating of FIBs is coordinated preferably based on the status and contents of the updated RIBs with respect to active CP blades.

**[0033]** When the PSM functionality of the router detects and determines a fatal fault in an active node of the router (block 506), the fault is localized and a continuous switchover is effectuated with respect to that node (either in the CP or DP domain) (block) 508). As a result, non-disruptive forwarding of the ingress packets or cells continues to take place in other data paths of the DP domain based on current update buffer and FIB conditions in the redundant data paths (block 510). Since the continuous switchover process takes place internally within the router (i.e., no control updates are generated towards the peer elements to which the router is coupled), the existing links involving the router continue to be maintained. Thus, even where a routing protocol requires that the underlying TCP/IP mechanism continuously provide status update information to the peers, there will be no teardown of the connections. In other words, there is no route flapping in the network due to fatal errors in the router's architecture, thereby greatly enhancing availability.

**[0034]** Upon completion of the switchover operation, a redundant CP node or blade becomes active, which condition is propagated throughout the router for recommencing the CP and DP update agent processes (i.e.,

resynchronization of the CP and DP domains) (block 508). Data forwarding processes continue uninterrupted based on the reconfigured update buffers and FIBs (block 512). The routing process methodology continues thereafter in normal manner using loosely-coupled, disjoint synchronization between the reconfigured DP/CP domains (block 514).

**[0035]** Based upon the foregoing Detailed Description, it should be readily apparent that the present invention advantageously provides a fault-tolerant routing architecture that is scalable and adaptable to any cluster-based or interconnection-based router design. By rendering the CP and DP domains disjoint and independently redundant, the effect of a fatal fault in one section of the router can be isolated from the remaining portions of the distributed design, which can continue to process the incoming data while the switchover process takes place.

**[0036]** It is believed that the operation and construction of the present invention will be apparent from the foregoing Detailed Description. While one or more of the exemplary embodiments of the invention shown and described have been characterized as being preferred, it should be readily understood that various changes and modifications could be made therein without departing from the scope of the present invention as set forth in the following claims.